The length of a leaf coloration on a random binary tree

by

A.M. Hamel and M.A. Steel
Department of Mathematics and Statistics
University of Canterbury, Christchurch, New Zealand.

No. 109 July, 1994

Keywords: binary tree, Fitch's algorithm, maximum parsimony tree, DNA/RNA sequences, probability.
AMS (MOS) Subject Classification: 05C05, 05A15, 92D20.

The length of a leaf coloration on a random binary tree

A.M. Hamel and M.A. Steel
Department of Mathematics and Statistics
University of Canterbury
Christchurch, New Zealand

July 11, 1994

Abstract

An assignment of colors to objects induces a natural integer weight on each tree that has these objects as leaves. This weight is called "parsimony length" in biostatistics, and is the basis of the "maximum parsimony" technique for reconstructing evolutionary trees. Equations for the average value (over all binary trees) of the parsimony length of both fixed and random colorations are derived using generating function techniques. This leads to asymptotic results that extend earlier results confined to just two colors. A potential application to DNA sequence analysis is outlined briefly.

1. Introduction

Let $\mathcal{B}(n)$, $n \geq 2$, denote the set of (unrooted) trees with n leaves (vertices of degree 1) labelled $1, 2, \ldots, n$, and with all remaining vertices unlabelled and of degree 3. Such trees, which we will simply call *binary trees*, are useful representations of evolutionary relationships in taxonomy. In that case, the set $[n] = \{1, 2, \ldots, n\}$ represents the extant taxa being classified, while the remaining vertices in the tree represent ancestral taxa. It is often convenient to represent the (global) ancestral taxon of all these taxa by a root vertex obtained by subdividing an edge (the "most ancient" edge) of the tree. Let $\mathcal{R}(n)$, $n > 1$, denote the set of all such edge rooted binary trees on leaf set $[n]$. We define $\mathcal{R}(1)$ as the singleton set consisting of an isolated (root) vertex labelled 1. Note for $n \geq 2$ the bijection:

$$\psi : \mathcal{B}(n) \to \mathcal{R}(n-1)$$

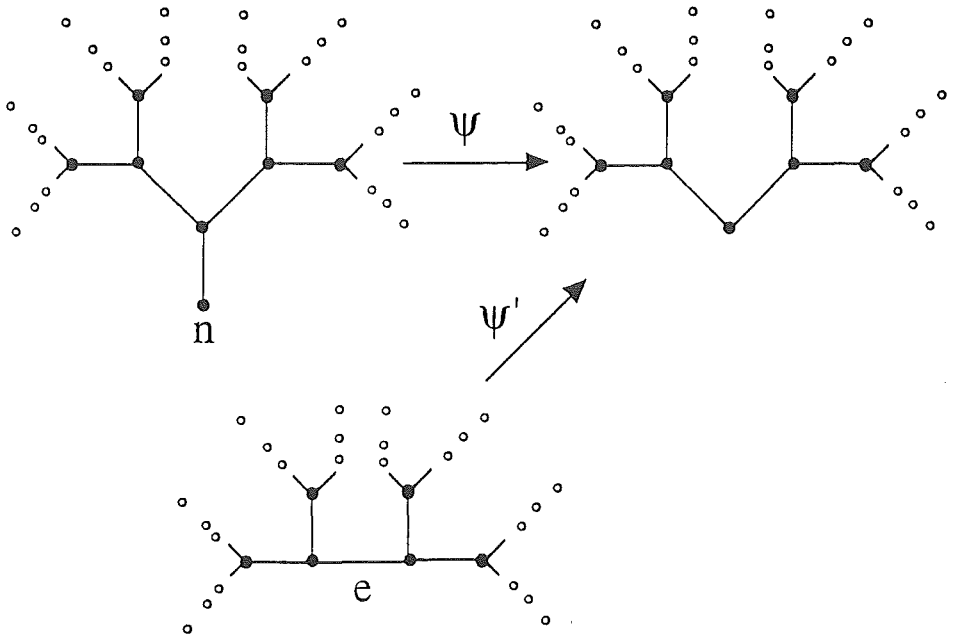


Figure 1: Bijections between rooted and unrooted binary trees.

where, if $T \in \mathcal{B}(n)$, $\psi(T)$ is the edge-rooted binary tree which results when leaf n and its incident edge are deleted, as in Figure 1. Edge subdivision also gives a bijection:

$$\psi' : \{(T, e) : T \in \mathcal{B}(n),\ e \in E(T)\} \to \mathcal{R}(n)$$

as in Figure 1. We let

$$B(n) := |\mathcal{B}(n)| \quad \text{and} \quad R(n) := |\mathcal{R}(n)|$$

for $n \geq 2$ and $n \geq 1$ respectively. Since $|E(T)| = 2n - 3$, for each $T \in \mathcal{B}(n)$, it follows (from ψ and ψ') that, for $n \geq 3$,

$$\begin{aligned} R(n) = (2n-3)B(n) = (2n-3)!! &= 3 \times 5 \times \cdots \times (2n-3) \\ &= \frac{(2n-2)!}{2^{n-1}(n-1)!}, \end{aligned} \tag{1}$$

a result dating back at least as far as 1870 to a paper by Schröder [9]. Thus, by applying Stirling's formula to $R(n)$,

$$\frac{R(n)}{n!} \sim \frac{1}{2\sqrt{\pi}}\, 2^n n^{-3/2}. \tag{2}$$

(A definition of "asymptotic" ($\sim$) appears at the beginning of Section 3).

Let χ be a coloration of $[n]$ by a set $\mathcal{C}$ of $r \geq 2$ colors. For example, in phylogenetic analysis each site j in a collection of n aligned DNA/RNA sequences (where $r = 2$ or 4) gives a coloration $\chi = \chi^j$ of $[n]$ for which $\chi^j(i)$ is the nucleotide that occurs at site j in the i-th sequence.

Given a tree T in $\mathcal{B}(n)$ or $\mathcal{R}(n)$ and a coloration χ of $[n]$ let $\ell(T, \chi)$ be the minimal number of edges of T that need to be assigned differently–colored ends in order to extend χ to a coloration of all the vertices of T (any such minimizing extension is called a *minimal extension* of χ for T). The number $\ell(T, \chi)$ is called

the *parsimony length* of χ on T, and it is the basis of the widely-used "maximum parsimony" technique for reconstructing evolutionary trees from aligned genetic sequences. This approach selects the tree(s) T which minimizes (minimize) the sum of $\ell(T, \chi^j)$ over all sites j in the sequences—this sum is the *length* of T on the sequences. Such a tree—a *maximum parsimony* tree—requires the fewest mutations to account for the variations in the aligned sequences.

The aim of this paper is to develop analytic techniques that would allow the length of the maximum parsimony tree on the original sequences to be compared with the average length of all binary trees on either (i) the original sequences or (ii) randomized versions of the original sequences (i.e. sequences generated randomly with the same expected frequencies of colors as the original sequences, as in Steel *et. al.* [11]). These two average values are obtained by applying respectively, functions μ_n and μ'_n (which we describe in Section 2) to each sequence site, and summing up the resulting values. An asymptotic formula for μ'_n is described in Section 3 and since, as we show, μ_n and μ'_n are asymptotically equivalent, this provides an asymptotic formula for μ_n as well. Our results exploit some special properties of the generating functions which count various classes of leaf labelled trees according to their parsimony length. In this sense the exact and asymptotic analyses compliment and extend the approaches of Carter *et. al.* [2], and Butler [1] respectively, both of which analysed similar systems of generating functions with just two colors (although the problems these authors considered were slightly different from ours).

First we describe a convenient technique for computing $\ell(T, \chi)$ known as the (first pass of) Fitch's algorithm (Fitch [3], Hartigan [5]). If $T' \in \mathcal{B}(n)$, subdivide any edge of T' to obtain a tree $T \in \mathcal{R}(n)$. Note that $\ell(T, \chi) = \ell(T', \chi)$. Now direct all edges of T away from the root and recursively assign nonempty subsets of colors to the vertices of T beginning with the leaves and progressing towards the root, as follows:

(1) leaf $i \in [n]$ is assigned the singleton set $\{\chi(i)\}$,

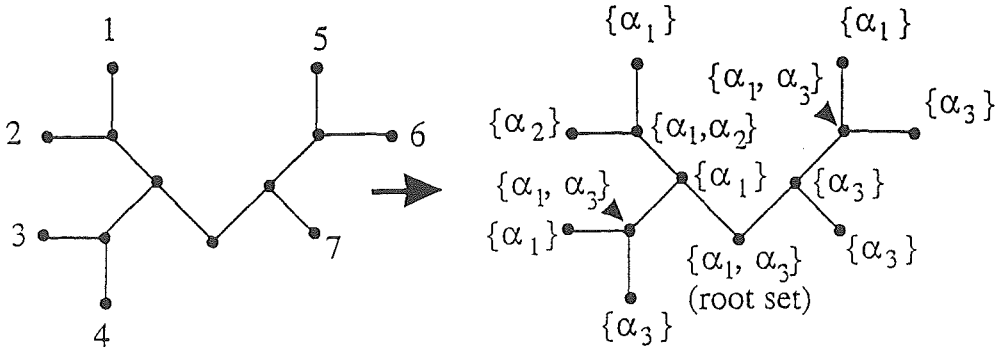
(2) once the descendants of vertex v have both been assigned sets A, B, then assign vertex v the set $A * B$, where $*$ is the (non-associative, binary) "parsimony operation" defined on $2^{\mathcal{C}} - \phi$,

$$A * B = \begin{cases} A \cap B, & \text{if } A \cap B \neq \phi \\ A \cup B, & \text{if } A \cap B = \phi. \end{cases}$$

The set assigned to the root of T is called the *root set* (In the case $T \in \mathcal{R}(1)$ the root set is just $\{\chi(1)\}$). These concepts are illustrated in Figure 2. A fundamental property of this procedure is the following:

Lemma 1 (Hartigan [5]) *$\ell(T, \chi)$ is the number of times an empty intersection (option 2 in the above description of $*$) is encountered in this assignment of sets of colors to the vertices of T. Furthermore the root set is precisely the set of those colors that occur in at least one minimal extension of χ for T.*

We will use both of these properties in Section 2.



$\chi(1) = \chi(3) = \chi(5) = \alpha_1$; $\chi(2) = \alpha_2$;
$\chi(4) = \chi(6) = \chi(7) = \alpha_3$.

Figure 2: A rooted tree with sets assigned to vertices by the parsimony operation.

Notation

(1) For convenience, we write

$$\underline{x} \quad \text{to denote} \quad (x_1, x_2, \ldots, x_r),$$

$$\underline{x}^{\underline{a}} \quad \text{to denote the monomial} \quad x_1^{a_1} x_2^{a_2} \ldots x_r^{a_r},$$

and

$$\underline{a}! \quad \text{to denote} \quad a_1! a_2! \ldots a_r! \; .$$

(2) We also write

$$[\underline{x}^{\underline{a}}] f(\underline{x}) \quad \text{to denote the coefficient of} \quad x_1^{a_1} x_2^{a_2} \ldots x_r^{a_r} \quad \text{in } f(\underline{x}),$$

as in Goulden and Jackson [4].

(3) $\mathcal{C} = \{\alpha_1, \ldots, \alpha_r\}$ will denote the set of colors which are assigned to the elements of the set $[n] = \{1, \ldots, n\}$. If $a_i = |\chi^{-1}(\alpha_i)|$, $i = 1 \ldots r$, we say χ is of *type* $\underline{a} = (a_1, \ldots, a_r)$. Thus, $a_i \geq 0$ and $\sum_{i=1}^{r} a_i = n$.

2. Calculations (Exact)

The aim of this paper is to calculate the two averages that we now define.

Definition 1 (μ_n **and** μ'_n) *Let $\mu_n(\underline{a})$ be the average, over all trees $T \in \mathcal{B}(n)$, of the length of a fixed coloration of $[n]$ of type $\underline{a}$ on T (by symmetry this is well-defined).*

For probability distribution $\underline{\phi} = (\phi_1, \phi_2, \ldots, \phi_r)$, $\phi_1 \geq 0, \phi_2 \geq 0, \ldots, \phi_r \geq 0$, $\sum_{i=1}^{r} \phi_i = 1$, let $\mu'_n(\underline{\phi})$ be the average, over all trees $T \in \mathcal{B}(n)$, of the expected length of a random coloration of $[n]$ on T. In this random coloration each element of $[n]$ is independently assigned color α_i with probability ϕ_i.

Note that $\mu'_n(\underline{\phi})$ is the average, over all trees $T \in \mathcal{B}(n)$, of

$$\sum_{\chi} \ell(T, \chi) \prod_{j=1}^{n} \phi_{\chi(j)} \; ,$$

and so

$$\mu'_n(\underline{\phi}) = \sum_{\underline{a}} \binom{n}{\underline{a}} \underline{\phi}^{\underline{a}}\, \mu_n(\underline{a})\ . \tag{3}$$

Here and elsewhere, a summation over $\underline{a}$ ranges over all nonnegative r–tuples $a_1, \ldots, a_r$ with $\sum_{i=1}^r a_i = n$. Also note that μ_n and μ'_n are symmetric functions in $a_1, \ldots, a_r$ and $\phi_1, \ldots, \phi_r$ respectively. The following generating function forms the basis for our calculations. For $\phi \neq A \subseteq C$, let $\quad T_A(\underline{x}, z) = \sum_{\underline{a},\ell} \dfrac{f_A(\underline{a},\ell)}{\underline{a}!}\, \underline{x}^{\underline{a}} z^{\ell}$, where $f_A(\underline{a},\ell)$ is the number of trees in $\mathcal{R}(n)$, $n \geq 1$, of parsimony length $\ell \geq 0$ and root set A for a fixed r–coloration of $[n]$ of type $\underline{a}$ (by symmetry considerations, $f_A(\underline{a},\ell)$ is well–defined). By Lemma 1, the set $\{T_A(\underline{x}, z),\ \emptyset \neq A \subseteq C\}$ satisfies the system of simultaneous quadratic equations described in Steel [1993],

$$T_A(\underline{x}, z) = \sum_{(B,C):B\cap C=A} \frac{1}{2} T_B(\underline{x}, z) T_C(\underline{x}, z) + \sum_{(B,C):\ {B\cap C=\emptyset \atop B\cup C=A}} \frac{z}{2} T_B(\underline{x}, z) T_C(\underline{x}, z) + \delta_A(\underline{x}) \tag{4}$$

where

$$\delta_A(\underline{x}) = \begin{cases} x_i & \text{if } A = \{\alpha_i\} \\ 0, & \text{if } |A| > 1. \end{cases} \tag{5}$$

For $r = 2$ this system can be treated by the multivariate Lagrange inversion formula (Goulden and Jackson [4]) to give an explicit closed–form expression for $f_A(\underline{a},\ell)$—see Carter *et. al.* [2], Steel [10].

Theorem 1 (Exact formulae) *Let* $T_A(\underline{x}) := T_A(\underline{x}, 1)$.

(i)

$$\mu_n(\underline{a}) = \frac{\underline{a}!}{R(n)}\, [\underline{x}^{\underline{a}}] \sum_{(A,B):A\cap B=\emptyset} \frac{1}{2}\, T_A(\underline{x}) T_B(\underline{x}) \left(1 - 2\sum_{i=1}^r x_i\right)^{-\frac{1}{2}}$$

(ii)

$$\mu'_n(\underline{\phi}) = \frac{n!}{R(n)} [x^n] \sum_{(A,B):A\cap B=\emptyset} \frac{1}{2}\, T_A(\underline{\phi} x) T_B(\underline{\phi} x)(1-2x)^{-\frac{1}{2}}$$

where $\underline{\phi} x = (\phi_1 x, \ldots, \phi_r x)$.

(iii)

$$\mu'_n(\underline{\phi}) = \mu'_{n-1}(\underline{\phi}) + \sum_{\substack{i,A:\\ \alpha_i \notin A}} \phi_i \frac{(n-1)!}{R(n-1)} [x^{n-1}] T_A(\underline{\phi} x).$$

Proof of Theorem 1 :

Let

$$\begin{aligned} R(\underline{x}, z) &:= \sum_{A\neq\emptyset} T_A(\underline{x}, z), \\ R(\underline{x}) &:= R(\underline{x}, 1). \end{aligned} \tag{6}$$

First observe that, from (4), we have the fundamental identity,

$$R(\underline{x},z) = \frac{1}{2}\,R^2(\underline{x},z) + (z-1)\sum_{\substack{(B,C):\\ B\cap C=\emptyset}} \frac{1}{2}\,T_B(\underline{x},z)T_C(\underline{x},z) + \sum_{i=1}^{r} x_i\ . \tag{7}$$

Putting $z=1$ in (7), we obtain

$$R(\underline{x}) = \frac{1}{2}R^2(\underline{x}) + \sum_{i=1}^{r} x_i \tag{8}$$

so that

$$R(\underline{x}) = 1 - \sqrt{1 - 2\sum_{i=1}^{r} x_i}\ . \tag{9}$$

In particular,

$$R(\underline{\phi}x) = 1 - \sqrt{1-2x} \tag{10}$$

and so

$$[x^n]R(\underline{\phi}x) = \frac{R(n)}{n!}. \tag{11}$$

Let

$$Q(\underline{x}) := \frac{\partial}{\partial z}_{|_{z=1}} R(\underline{x},z). \tag{12}$$

Then, from (7),

$$Q(\underline{x}) = Q(\underline{x})R(\underline{x}) + \sum_{\substack{(B,C):\\ B\cap C=\emptyset}} \frac{1}{2}\,T_B(\underline{x})T_C(\underline{x})$$

hence

$$Q(\underline{x}) = \sum_{(B,C):B\cap C=\emptyset} \frac{1}{2}\,T_B(\underline{x})T_C(\underline{x})\cdot(1-R(\underline{x}))^{-1}\ .$$

Applying (9) gives

$$Q(\underline{x}) = \sum_{(B,C):B\cap C=\emptyset} \frac{1}{2}\,T_B(\underline{x})T_C(\underline{x})\left(1-2\sum_{i=1}^{r} x_i\right)^{-\frac{1}{2}}\ . \tag{13}$$

Now from (12), $\underline{a}![\underline{x}^{\underline{a}}]Q(\underline{x})$ is the sum $\sum_{\ell} \ell f(\underline{a},\ell)$ where $f(\underline{a},\ell) = \sum_{A\neq\emptyset} f_A(\underline{a},\ell)$, the total number of trees $T\in\mathcal{R}(n)$ of length ℓ for a coloration χ of $[n]$ of type $\underline{a}$. Thus $\frac{\underline{a}!}{R(n)}[\underline{x}^{\underline{a}}]Q(\underline{x})$ is the average length over all trees in $\mathcal{R}(n)$ of the length of χ. However, each edge rooting of a binary tree leads to an identical parsimony length (i.e. the position of the root is irrelevant to the length) so this quantity is also the average length over all trees in $\mathcal{B}(n)$ of the length of χ, which in view of (13) establishes part (i).

(ii) Applying part (i) to (3) we obtain

$$\mu_n'(\underline{\phi}) = \sum_{\underline{a}} \binom{n}{\underline{a}} \underline{\phi}^{\underline{a}} \frac{\underline{a}!}{R(n)}\,[\underline{x}^{\underline{a}}]F(\underline{x}) \tag{14}$$

where $F(\underline{x}) = \sum_{(A,B):A\cap B=\emptyset} \frac{1}{2} T_A(\underline{x}) T_B(\underline{x}) (1 - 2\sum_{i=1}^{r} x_i)^{-\frac{1}{2}}$.

Rewriting (14), we have,

$$\mu'_n(\underline{\phi}) = \frac{n!}{R(n)} \sum_{\underline{a}} \underline{\phi}^{\underline{a}} [\underline{x}^{\underline{a}}] F(\underline{x}) = \frac{n!}{R(n)} [x^n] F(\underline{\phi} x)$$

as required.

(*iii*) For any tree $T' \in \mathcal{R}(m)$, with root vertex ρ and subject to a random coloration χ of $[m]$ according to $\underline{\phi}$, let $S(T')$ denote the (random variable) root set of T' (as defined in Section 1).

By the bijection $\psi : \mathcal{B}(n) \to \mathcal{R}(n-1)$ and Lemma 1 (rooting T on the edge incident with the leaf labelled n), we have,

$$\ell(T,\chi) = \ell(\psi(T),\chi') + \delta(T,\chi) \tag{15}$$

$$\text{where} \quad \delta(T,\chi) = \begin{cases} 1 & \text{if } \chi(n) \notin S(\psi(T)) \\ 0 & \text{otherwise} \end{cases},$$

and where χ' is the restriction of χ to $[n-1]$.

For T in $\mathcal{B}(n)$ or $\mathcal{R}(n-1)$, let $\mu(T)$ denote the expected value of $\ell(T,\chi)$ for a random χ (generated according to $\underline{\phi}$). Then from (15) we have

$$\mu(T) = \mu(\psi(T)) + \text{Prob}[\delta(T,\chi) = 1] \,. \tag{16}$$

Now,

$$\begin{aligned} \text{Prob}[\delta(T,\chi) = 1] &= \text{Prob}[\chi(n) \notin S(\psi(T))] \\ &= \sum_{\substack{i,A: \\ \alpha_i \notin A}} \phi_i \text{Prob}[S(\psi(T)) = A] \,. \end{aligned} \tag{17}$$

Also, by definition,

$$\mu'_n(\underline{\phi}) = \frac{1}{B(n)} \sum_{T\in\mathcal{B}(n)} \mu(T), \tag{18}$$

while

$$\begin{aligned} \mu'_{n-1}(\phi) = \frac{1}{B(n-1)} \sum_{T'\in\mathcal{B}(n-1)} \mu(T') &= \frac{1}{B(n)} \sum_{T'\in\mathcal{B}(n-1)} (2n-3)\mu(T') \\ &= \frac{1}{B(n)} \sum_{T\in\mathcal{B}(n)} \mu(\psi(T)) \,. \end{aligned} \tag{19}$$

Thus, combining (16)–(19) we have

$$\mu'_n(\phi) = \mu'_{n-1}(\underline{\phi}) + \sum_{\substack{i,A: \\ \alpha_i \notin A}} \phi_i \frac{1}{B(n)} \sum_{T\in\mathcal{B}(n)} \text{Prob}[S(\psi(T)) = A] \,. \tag{20}$$

Now,

$$\frac{1}{B(n)} \sum_{T\in\mathcal{B}(n)} \text{Prob}[S(\psi(T)) = A] = \frac{1}{R(n-1)} \sum_{T'\in\mathcal{R}(n-1)} \text{Prob}[S(T') = A] \,. \tag{21}$$

Also, $n![x^n]T_A(\underline{\phi}x) = \sum_{\underline{a}} \binom{n}{\underline{a}} \underline{\phi}^{\underline{a}} f_A(\underline{a})$, and so,

$$\begin{aligned} n![x^n]T_A(\underline{\phi}x) &= \sum_{\underline{a}} \underline{\phi}^{\underline{a}} \sum_{\chi:\chi \text{ has } \atop \text{type } \underline{a}} \sum_{T\in\mathcal{R}(n) \atop S(T,\chi)=A} 1 \\ &= \sum_{T\in\mathcal{R}(n)} \sum_{\underline{a}} \sum_{\chi:\chi \text{ has type } \underline{a} \atop \text{and } S(T,\chi)=A} \underline{\phi}^{\underline{a}} \\ &= \sum_{T\in\mathcal{R}(n)} \sum_{\chi:S(T,\chi)=A} \text{Prob}[\chi] \\ &= \sum_{T\in\mathcal{R}(n)} \text{Prob}[S(T,\chi)=A] \,. \end{aligned}$$

Thus, the term on the right of (21) is just

$$\frac{(n-1)!}{R(n-1)}[x^{n-1}]T_A(\underline{\phi}x)$$

which, together with (20) establishes part (*iii*), thereby completing the proof of Theorem 1.

3. Calculations (asymptotic)

In this section we obtain asymptotic results concerning $\mu'_n(\underline{\phi})$ and $\mu_n(\underline{a})$. Theorem 2 below shows that $\mu'_n(\underline{\phi})$ and $\mu_n(\underline{a})$ are asymptotically equivalent since they both grow linearly with n, and their difference (when $\underline{\phi} = \frac{1}{n}\underline{a}$) is bounded by a term of order $n^{\frac{1}{2}}$. The theorem also provides a prescription for calculating, in principle, their asymptotic values by solving a system of simultaneous quadratic equations involving real numbers. In the case of two colors this can be done analytically, but generally numerical techniques would seem to be required. However, in the case of equifrequency colorations ($\phi_i = \frac{1}{r}$) the resulting system is considerably simpler, being of dimension r rather than $2^r - 1$, and we solve this for $r \leq 4$ in Corollary 1. A second corollary provides a biologically–oriented application.

We adopt the standard notation $f(n) \sim g(n)$ if $\lim_{n\to\infty} \frac{f(n)}{g(n)} = 1$, and $f(n) = O(g(n))$ if $\frac{f(n)}{g(n)}$ is bounded as $n \to \infty$.

Theorem 2 (Asymptotic formulae) *(i) $\mu'_n(\underline{\phi}) \sim \mu' n$ where $\mu' = \mu'(\underline{\phi})$ is given by*

$$\mu' = \sum_{(A,B):A\cap B=\emptyset} t_A t_B,$$

and where the numbers $t_A = T_A(\underline{x})|_{\underline{x}=\frac{1}{2}\underline{\phi}}, \emptyset \neq A \in \mathcal{C}$ satisfy the simultaneous system:

$$t_A = \sum_{(B,C):B*C=A} \frac{1}{2} t_B t_C + \delta_A(\underline{\phi})$$

with δ_A given by (5).

(ii) For $r = 2$ colors,

$$\mu' = \frac{2}{3}\left(1 - \sqrt{1 - 3\phi_1\phi_2}\right).$$

(iii) $\mu_n(\underline{a}) \sim n\mu'(\underline{\phi})$ for $\underline{\phi} = \frac{1}{n}\underline{a}$. Indeed, $|\mu_n(\underline{a}) - \mu'_n(\underline{\phi})| \leq \sqrt{n(r-1)}/2$ for all n.

Proof of Theorem 2:

(*i*) We first recall a special case of Lemma 1(i) of Meir, Moon and Mycielski [6]: suppose $F(x)$ and $G(x)$ are power series in x, and that

$$[x^n]F(x) = O(\rho^{-n}n^{-\frac{3}{2}})$$

$$[x^n]G(x) \sim b\rho^{-n}n^{-\frac{1}{2}}.$$

and $F(\rho) \neq 0$. Then

$$[x^n]F(x)G(x) \sim F(\rho)[x^n]G(x). \tag{22}$$

Taking $G(x) = (1-2x)^{-\frac{1}{2}}$ we have, from (2), that

$$[x^n]G(x) = \frac{R(n+1)}{n!} = (2n-1)\frac{R(n)}{n!} \sim \frac{1}{\sqrt{\pi}}\rho^{-n}n^{-\frac{1}{2}}, \tag{23}$$

where $\rho = \frac{1}{2}$. Now take $F(x) = T_A(\underline{\phi}x)T_B(\underline{\phi}x)$ for any pair of nonempty sets $A, B \subseteq \mathcal{C}$. Since for all $C \subseteq \mathcal{C}$, $C \neq \emptyset$ the power series $T_C(\underline{\phi}x)$ has all nonnegative coefficients, we have:

$$\begin{aligned}
|[x^n]F(x)| &= |[x^n]T_A(\underline{\phi}x)T_B(\underline{\phi}x)| \\
&\leq [x^n]\left(\sum_{C\neq\emptyset} T_C(\underline{\phi}x)\right)^2 \\
&= [x^n]R(\underline{\phi}x)^2 \quad \text{from (6)} \\
&= [x^n](2R(\underline{\phi}x) - 2x) \quad \text{from (8)} \\
&= 2\frac{R(n)}{n!} \text{ from (11)} \\
&\sim \frac{1}{\sqrt{\pi}}\rho^{-n}n^{-\frac{3}{2}}
\end{aligned}$$

for $\rho = \frac{1}{2}$, from (2).

Thus $[x^n]F(x) = O(\rho^{-n}n^{-\frac{3}{2}})$, and so, from (23) and the fact that F has nonnegative coefficients (so that $F(\rho) \neq 0$), we can apply (22) to Theorem 1 (*ii*) to deduce that

$$\begin{aligned}
\mu'_n &= \frac{n!}{R(n)} \sum_{(A,B):A\cap B=\emptyset} \frac{1}{2}T_A(\underline{\phi}x)T_B(\underline{\phi}x)|_{x=\frac{1}{2}}[x^n](1-2x)^{-\frac{1}{2}} \\
&\sim n \sum_{(A,B):A\cap B=\emptyset} T_A\left(\frac{1}{2}\underline{\phi}\right)T_B\left(\frac{1}{2}\underline{\phi}\right),
\end{aligned}$$

as claimed.

The prescribed system for $t_A := T_A(\frac{1}{2}\underline{\phi})$ follows from (4) by putting $z = 1$ and $\underline{x} = \underline{\phi}x$.

(*ii*) This result follows from part (*iii*) of Theorem 2, and the analogous result for $\mu_n(\underline{a})$ from Moon and Steel [7]. However it can also be derived more directly from Theorem 2 (*i*). We have, for $\mathcal{C} = \{\alpha, \beta\}$,

$$\mu' = 2T_{\{\alpha\}}T_{\{\beta\}}, \tag{24}$$

where $T_{\{\alpha\}} = T_{\{\alpha\}}(\frac{1}{2}\underline{\phi})$, $T_{\{\beta\}} = T_{\{\beta\}}(\frac{1}{2}\underline{\phi})$ and $T_{\{\alpha,\beta\}} = T_{\{\alpha,\beta\}}(\frac{1}{2}\underline{\phi})$ satisfy the system:

$$\begin{aligned} T_{\{\alpha\}} &= \frac{1}{2}T_{\{\alpha\}}^2 + T_{\{\alpha\}}T_{\{\alpha,\beta\}} + \frac{\phi_1}{2} \\ T_{\{\beta\}} &= \frac{1}{2}T_{\{\beta\}}^2 + T_{\{\beta\}}T_{\{\alpha,\beta\}} + \frac{\phi_2}{2} \\ T_{\{\alpha,\beta\}} &= \frac{1}{2}T_{\{\alpha,\beta\}}^2 + T_{\{\alpha\}}T_{\{\beta\}} \end{aligned}$$

Butler [1] solved this system, and from his equation (26), we have:

$$\begin{aligned} T_{\{\alpha\}}^2 &= \frac{1}{3}(-2 + 3\phi_1 + 2\sqrt{P}) \\ T_{\{\beta\}}^2 &= \frac{1}{3}(1 - 3\phi_1 + 2\sqrt{P}) \end{aligned}$$

where $P = 1 - 3\phi_1\phi_2$ and from this we can obtain μ' directly from (24).

(*iii*) We first claim that,

$$|\mu_n(\underline{a}) - \mu_n(\underline{a}')| \leq \frac{1}{2}|\underline{a} - \underline{a}'|_1, \tag{25}$$

where $|\ |_1$ denotes the l_1 norm on $\mathbf{R}^r$. Since the components of $\underline{a}$ and $\underline{a}'$ both sum to n, $|\underline{a} - \underline{a}'|_1 = 2k$ for some integer k. In that case we can find two colorations χ, χ' of $[n]$ of types $\underline{a}$, $\underline{a}'$ respectively, and such that χ and χ' agree on all but k elements of $[n]$.

Now, for any binary tree T, it is easily checked that $\ell(T, \chi') \leq \ell(T, \chi) + k$ since any mimimal extension of χ for T produces an extension χ'' of T by just changing the colors of the (at most k) leaves of T for which χ and χ' disagree (and thereby increasing the number of edges of T with differently colored ends by at most k). Although χ'' may not be a minimal extension of χ' for T, we nevertheless obtain the claimed inequality. Conversely, $\ell(T, \chi) \leq \ell(T, \chi') + k$, and so

$$|\ell(T, \chi) - \ell(T, \chi')| \leq k$$

which, upon averaging over all binary trees, gives,

$$|\mu_n(\underline{a}) - \mu_n(\underline{a}')| \leq k,$$

which establishes (25).

Now from (3),

$$\mu'_n = E[\mu_n(\underline{A})],$$

the expected value of $\mu_n(\underline{A})$ where $\underline{A}$ is drawn from a multinomial distribution with parameters n and $\phi_1, \ldots, \phi_r$.

Then,

$$\begin{aligned} |\mu_n(\underline{a}) - \mu'_n| &= |E[\mu_n(\underline{a}) - \mu_n(\underline{A})]| \\ &\leq E[|\mu_n(\underline{a}) - \mu_n(\underline{A})|] \\ &\leq \frac{1}{2} E[|\underline{a} - \underline{A}|_1], \text{ from (25)} \\ &= \frac{1}{2} \sum_{i=1}^{r} E[|a_i - A_i|]. \end{aligned} \tag{26}$$

Now A_i has a binomial distribution with parameters n and ϕ_i, and since

$$E[A_i] = n\phi_i = a_i,$$

we have, applying the convex version of Jensen's inequality (Rényi [8]):

$$\begin{aligned} E[|a_i - A_i|] &\leq \sqrt{E[(a_i - A_i)^2]} \\ &= \sqrt{Var[A_i]} \\ &= \sqrt{n\phi_i(1-\phi_i)}, \end{aligned}$$

so that, from (26), $|\mu_n(\underline{a}) - \mu'_n(\underline{\phi})| \leq \frac{1}{2} \sum_{i=1}^{r} \sqrt{n\phi_i(1-\phi_i)}$, which, by the concave version of Jensen's inequality, is at most $\sqrt{n(r-1)}/2$.

Corollary 1 (Equifrequency Colorations) *Suppose $\phi_i = \frac{1}{r}$ for $i = 1, \ldots, r$. Then*

$$\mu' = \sum_{(j,k): j+k \leq r;\, j,k \geq 1} \frac{r!}{j!k!(r-j-k)!} t_j t_k$$

where the t_i satisfy the system:

$$t_i = \sum_{(j,k)} \frac{1}{2} \pi_{ijk} t_j t_k + \delta_{i,1} \frac{1}{2r},$$

for $i = 1, \ldots, r$ and where $\delta_{i,1} = 1$ if $i = 1$ and 0 otherwise, and where π_{ijk} is the number of sets of sizes j and k which, under the parsimony operation (*), give a specific root set of size i for the tree, i.e.,

$$\pi_{ijk} = \begin{cases} \binom{i}{j}, & \text{if } j + k = i \\ \binom{r-i}{j-i}\binom{r-j}{k-i}, & \text{if } j, k \geq i \\ 0, & \text{else} \end{cases}$$

Examples: For $r = 2$ we have

$$\begin{aligned} t_1 &= \frac{1}{2}t_1^2 + t_1t_2 + \frac{1}{4} \\ t_2 &= \frac{1}{2}t_2^2 + t_1^2 \end{aligned}$$

which gives $t_1 = \frac{1}{\sqrt{6}}$, $t_2 = 1 - \frac{2}{\sqrt{6}}$ and so from Corollary 1 we obtain $\mu' = \frac{1}{3}$, which agrees with Theorem 2 (ii). For $r > 2$ it seems necessary to solve the system $\{t_i\}$ by numerical methods. For $r = 3$ and 4, the equations become

$$\begin{aligned} t_1 &= \frac{1}{2}t_1^2 + 2t_1t_2 + t_1t_3 + t_2^2 + \frac{1}{6} \\ t_2 &= t_1^2 + \frac{1}{2}t_2^2 + t_2t_3 \\ t_3 &= 3t_1t_2 + \frac{1}{2}t_3^2 \end{aligned}$$

and

$$\begin{aligned} t_1 &= \frac{1}{2}t_1^2 + 3t_1t_2 + 3t_2^2 + 3t_1t_3 + 3t_2t_3 + t_1t_4 + \frac{1}{8} \\ t_2 &= t_1^2 + \frac{1}{2}t_2^2 + 2t_2t_3 + t_3^2 + t_2t_4 \\ t_3 &= 3t_1t_2 + \frac{1}{2}t_3^2 + t_3t_4 \\ t_4 &= 4t_1t_3 + 3t_2^2 + \frac{1}{2}t_4^2 \end{aligned}$$

respectively, and we find that $t = (0.24855, 0.06755, 0.051705)$ and $\mu' = 0.4714$ for $r = 3$ and $t = (0.17656, 0.0339, 0.01843, 0.01660)$ and $\mu' = 0.5507$ for $r = 4$.

As a second, and biologically–oriented application of Theorem 2, let us, as in Section 1, regard a collection of n aligned DNA sequences of length c as a collection $\chi^1, \ldots \chi^c$ of r–colorations of $[n]$ (for $r = 2$ or 4). Let $\ell(T)$ denote the length of $T \in \mathcal{B}(n)$ for this data; that is,

$$\ell(T) = \sum_{j=1}^{c} \ell(\chi^j, T)$$

and let $\bar{\ell}$ be the average value of $\ell(T)$ over $\mathcal{B}(n)$. We also consider a randomized version of $\bar{\ell}$ as follows. Let $\ell^*(T)$ be the expected length of a given binary tree T on sequences randomly generated by assigning each of the c sites in sequence i a color α_j with probability ϕ_j^i, as in Steel *et. al.* [11]. Let $\bar{\ell}^*$ denote the average value of $\ell^*(T)$ over $B(n)$. Finally, let $\mathcal{P}(n)$ denote the set of partitions of n into at most r–parts (thus $\mathcal{P}(n) = \{(p_1, \ldots, p_r) : p_1 \geq p_2 \geq \ldots \geq p_r \geq 0, \sum_{i=1}^{r} p_i = n\}$).

Corollary 2 *Asymptotically (as $n \to \infty$),*

$$(i) \quad \bar{\ell} \sim n \sum_{\underline{p} \in \mathcal{P}(n): N(\underline{p}) > 0} \mu' \left(\frac{1}{n}\underline{p}\right) N\left(\underline{p}\right)$$

where $N(\underline{p})$ is the number of sites j for which the type of χ^j, arranged in decreasing order gives partition $\underline{p}$.

$$(ii) \quad \overline{\ell}^* \sim cn\mu'(\underline{\phi}), \quad \textit{where} \quad \underline{\phi} = \frac{1}{n}\sum_{i=1}^{n} \underline{\phi}^i,$$

Proof of Corollary 2: (i)

$$\begin{aligned}
\overline{\ell} &= \frac{1}{B(n)} \sum_{T \in \mathcal{B}(n)} \ell(T) \\
&= \frac{1}{B(n)} \sum_{T \in \mathcal{B}(n)} \sum_{j=1}^{c} \ell(\chi^j, T) \\
&= \sum_{j=1}^{c} \frac{1}{B(n)} \sum_{T \in \mathcal{B}(n)} \ell(\chi^j, T) \\
&= \sum_{j=1}^{c} \mu_n(\underline{a}^{(j)})
\end{aligned}$$

where $\underline{a}^j$ is the type of χ^j. Thus

$$\overline{\ell} = n \sum_{\underline{p} \in \mathcal{P}(n): N(\underline{p}) > 0} \mu_n(\underline{p}) N(\underline{p})$$

and the result now follows from parts (i) and (iii) of Theorem 2.

(ii)

$$\overline{\ell}^* = cE'[\mu_n(\underline{A})] \tag{27}$$

where E' denotes expectation at a single site in the probability space described above. Since A_j is a sum of n independent (but not necessarily identical) $0-1$ random variables, D_{ij}, with $\text{Prob}[D_{ij} = 1] = \phi_j^i$, we have that

$$\frac{1}{n}\underline{A} \to_p \underline{\phi} \tag{28}$$

where $\to_p$ denotes convergence in probability (as $n \to \infty$) and

$$\underline{\phi} = \frac{1}{n}\sum_{i=1}^{n} \underline{\phi}^i$$

Now from (3), (25), and (28), it can be checked that

$$\left| \frac{1}{n}\mu_n'\left(\frac{1}{n}\underline{A}\right) - \frac{1}{n}\mu_n'(\underline{\phi}) \right| \to_p 0 \tag{29}$$

as $n \to \infty$.

Also, by Theorem 2 parts (i) and (iii) respectively,

$$\left| \frac{1}{n}\mu_n'(\underline{\phi}) - \mu'(\underline{\phi}) \right| \to 0,$$

$$\left|\frac{\mu_n(\underline{A})}{n} - \frac{1}{n}\mu_n'\left(\frac{1}{n}\underline{A}\right)\right| \to_p 0$$

as $n \to \infty$, thus

$$\frac{\mu_n(\underline{A})}{n} \to_p \mu'(\underline{\phi}) \tag{30}$$

as $n \to \infty$.

Now, from (27),

$$\overline{\ell}^* = cnE'\left[\frac{\mu_n(\underline{A})}{n}\right]$$

which together with (30) establishes part (i).

Acknowledgements

The first author acknowledges the support of a postdoctoral fellowship from the Natural Sciences and Engineering Research Council of Canada.

References

[1] J.P. BUTLER, *Fractions of trees with given root traits; the limit of large trees*, J. Theor. Biol., 147 (1990), pp. 265–274.

[2] M. CARTER, M.D. HENDY, D. PENNY, L.A. SZÉKELY, AND N.C. WORMALD, *On the distribution of lengths of evolutionary trees*, SIAM J. Discrete Math., 3 (1990), pp. 38–47.

[3] W.M. FITCH, *Towards defining the course of evolution: Minimum change for a specific tree topology*, Syst. Zool., 20 (1971), pp. 406–416.

[4] I.P. GOULDEN AND D.M. JACKSON, *Combinatorial Enumeration*, Wiley, New York, 1983.

[5] J.A. HARTIGAN, *Minimum mutation fits to a given tree*, Biometrics, 29 (1973), pp. 53–65.

[6] A. MEIR, J.W. MOON AND J. MYCIELSKI, *Hereditary finite sets and identity trees*, J. Comb. Theory B, 35 (1983), pp. 142–155.

[7] J.W. MOON AND M.A. STEEL, *A limiting theorem for parsimoniously bicoloured trees*, Appl. Math. Lett., 6 (1993), pp. 5–8.

[8] A. RÉNYI, *Probability Theory*, North Holland, Amsterdam, 1970.

[9] E. SCHRÖDER, *Vier combinatorische Probleme*, Zeitschrift für Mathematik und Physik, 15 (1870), pp. 361–376.

[10] M.A. STEEL, *Decompositions of leaf-colored binary trees*, Adv. in Appl. Math., 14 (1993), pp. 1–24.

[11] M.A. STEEL, P.J. LOCKHART, AND D. PENNY, *Confidence in evolutionary trees from biological sequence data*, Nature, 364 (1993), pp. 440–442.